VI CONGRESO IBÉRICO de AgroIngeniería
5 a 7 de Septiembre | 2011
Universidade de Évora | Portugal

Consecución de modelos robustos en una aplicación NIRS industrial mediante MSPC sobre espacios PLS

Adolfo Moya-González[1], Pilar Barreiro[1], Belén Diezma[1], Jaime Ortiz-Cañavate[1].
[1] Universidad Politécnica de Madrid. LPF-TAGRALIA. Avda. Complutense s/n, 28040 Madrid, España. E-mail: adolfo.moya@upm.es

Resumen

El presente trabajo presenta los resultados obtenidos mediante la aplicación de modelos PLS a una base de datos de espectros NIR de cebolla obtenidos de una aplicación en línea implementada en la industria. Especialmente se centra en los efectos de la selección de individuos basada en estadísticos multivariantes calculados a partir de los modelos (T2 de Hotelling) y en el pre-procesado de los espectros para la corrección de las fuentes de variación no deseadas. Los resultados obtenidos muestran cómo la selección de individuos en cuanto a sus valores de T^2, combinada con el pre-procesado de los espectros, permite general un modelo PLS más estable debido al menor número de variables latentes implicadas. El mencionado modelo presenta errores típicos de estimación similares al resto de los modelos estudiados, tanto para los datos empleados en la calibración, como en la validación. Sin embargo sí se produce una mejora en el coeficiente de determinación cuando el modelo es validado frente a nuevos datos no empleados en su generación. Nuevas técnicas para alcanzar un mejor conocimiento de las fuentes de variación implicadas deberán ser aplicadas.

Abstract

The present study presents the results obtained by PLS modeling application on a NIR spectra onion database obtained from an on-line application implemented at the industry. The work is specially centered on the effects of sample selection based on multivariate statistical process control statistics (Hotelling T^2) calculated from the PLS space generated and the effect of the preprocessing methods for undesirable variability minimization. The results obtained show how the sample selection by its Hotelling T^2 values, combined with the spectra preprocessing allows the generation of a more parsimonious PLS model. The aforementioned model standard error of estimation is similar to the other models tested, both for calibration and validation data. Nevertheless, there is an improvement for the determination coefficient when the model is validated with new data not considered for its calibration. New techniques regarding a better knowledge of the implied sources of variation have to be tested.

Palabras Clave: robustez, espectroscopía NIR, cebolla, clasificación, variable latente
Keywords: robustness, NIR spectroscopy, onion, classification, latent variable

1. INTRODUCCIÓN Y OBJETIVOS

La aplicación de espectroscopía en el infrarrojo cercano (NIRS) que nos ocupa se desarrolló para la determinación del contenido en sólidos solubles (SSC) en cebollas, durante los años 2001 a 2003 (empleando medidas off-line) y fue automatizada y transferida a la industria donde opera desde la campaña 2004 (medidas on-line) hasta la actualidad. A partir de una base de datos de espectros off-line empleados para la calibración del modelo de regresión lineal (MLR) se definió en 2004 un espacio de componentes principales (PCA) capaz de identificar medidas anómalas (Barreiro, Henche et al. 2004).

La espectroscopía NIR es especialmente sensible a los efectos de parámetros interferentes. Considerando que el caso abordado en este trabajo es una aplicación NIR industrial aplicada a cebollas provenientes de un programa de mejora, la variabilidad interferente es muy elevada. Esta variabilidad es debida tanto a las variaciones estacionales como entre campañas, tanto de los equipos como de los bulbos a seleccionar. Por este motivo resulta imprescindible verificar el funcionamiento de los modelos a lo largo del tiempo (Fig1).

La aplicación del análisis multivariante de control de procesos (MSPC) sobre el mencionado PCA permite en la actualidad identificar anomalías en el funcionamiento del equipo así como individuos extraños (Barreiro, Ruiz-Altisent et al. 2005). Al operar el equipo en condiciones industriales, las fuentes de variación, a las que los modelos NIRS son especialmente sensibles, son difícilmente controlables. El estudio de varias técnicas de modelización (Barreiro, Chauchard et al. 2005) y de los efectos de un conocido factor de interferencia como es la temperatura (Barreiro, Moya-González et al. 2005) ha demostrado que los modelos MLR empleados, a pesar de ofrecer precisiones moderadas, resultan más estables ante nuevas fuentes de variación que otras técnicas de ajuste como la regresión de mínimos cuadrados parciales (PLS).

Una reflexión sobre el funcionamiento del sistema

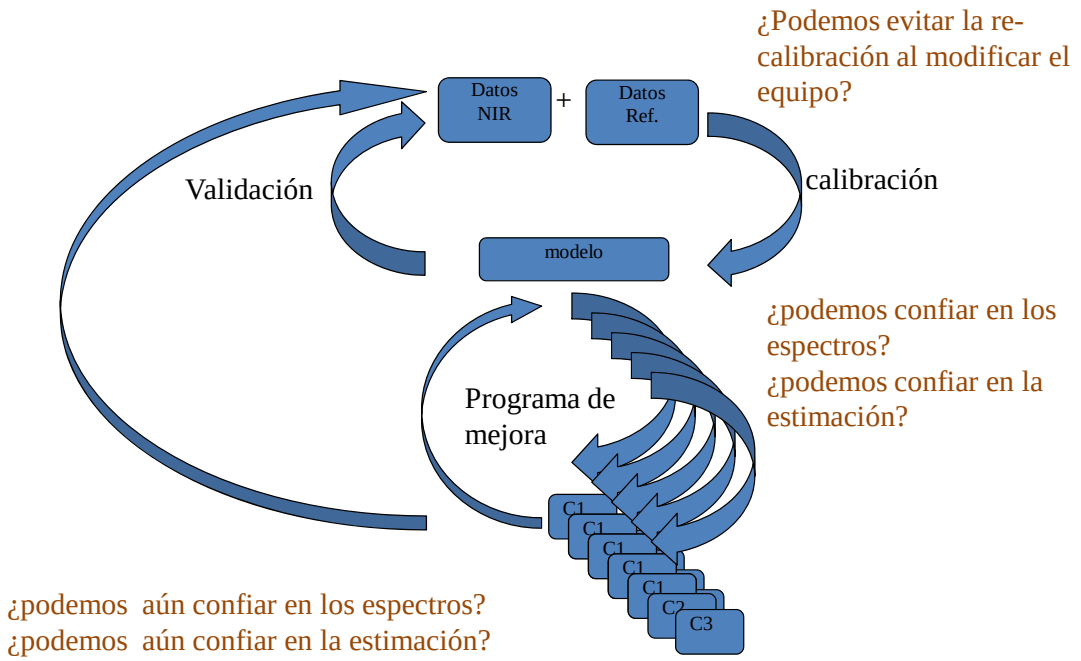

Figura 1. *Las fuentes de variación interferente y la evolución del material vegetal implican la necesidad de verificar y reajustar el funcionamiento de los modelos de estimación*

El empleo de métodos de pre-procesado de espectros permite la reducción de los efectos debidos a las nuevas fuentes de variación interferentes. Esta mejora de los espectros resulta de utilidad para la transferencia de calibración (CT) entre campañas (Moya-González, Barreiro et al. 2006; Moya-González, Barreiro et al. 2010). Adicionalmente, los espectros pre-procesados pueden ser empleados para la identificación de espectros anómalos mediante estadísticos MSPC sobre espacios PCA definidos a partir de una base de datos de espectros on-line (Moya-González, Barreiro et al. 2009; Ortiz-Cañavate, Moya-González et al. 2010).

El pre-procesado de los espectros, su proyección sobre un espacio PCA y aplicación del MSPC, permite una identificación de los espectros anómalos muy sensible a las nuevas

fuentes de variación incorporadas. Al tratarse de un sistema de clasificación que opera sobre material vegetal en constante evolución y en un entorno industrial, la gran cantidad de nuevas fuentes de variación incorporada de año en año provoca la identificación de espectros anómalos debida a parámetros que no afectan necesariamente a la estimación del SSC. Por lo tanto, el sistema de diagnóstico de operación desarrollado hasta la fecha, se centra más en el funcionamiento del equipo y no específicamente en el rendimiento del modelo.

En el presente trabajo se aborda la aplicación de técnicas estadísticas MSPC aplicadas sobre espacios PLS para diagnosticar los efectos de las nuevas fuentes de variación sobre los modelos de estimación del SSC en cebolla mediante NIRS así como el efecto del pre-procesado de los espectros originales sobre los modelos de estimación.

2. MATERIAL Y MÉTODOS

En este estudio se aplicarán métodos para la selección de las bases de datos de calibración, que combinadas con diversos métodos de pre-procesado de los espectros para la corrección de la varianza interferente permiten la creación de modelos PLS más robustos (Zeaiter, Rutledge et al. 2009). Se espera que la aplicación de técnicas MSPC sobre los espacios de variables latentes definidos mediante PLS permita una identificación más operativa de los espectros anómalos, ya que los estadísticos multivariantes, se definirán sobre el espacio de las variables latentes y por lo tanto serán más sensibles a las fuentes de variación que afectan al rendimiento del modelo de estimación. Con esto se pretende la consecución de un sistema de diagnóstico del funcionamiento del modelo.

La base de datos empleada para la calibración de los modelos PLS comprende 2697 espectros NIR, obtenidos durante las campañas 2001 y 2002 (medidas off-line) y la campaña 2004 (medidas on-line). Para la validación de los modelos se emplea una base de datos de 76 espectros NIR correspondientes a la campaña 2008 (medidas on-line), última campaña para la que los espectros están disponibles y que por lo tanto recoge las variaciones acumuladas, tanto por el sistema como por la variedad de cebolla.

Los modelos PLS se ajustaron sobre los espectros brutos (sin pre-procesado) y posteriormente fueron validados para los datos de la campaña 2008. La identificación de individuos anómalos se llevó a cabo mediante la determinación del estadístico T^2 de Hotelling sobre el espacio PLS generado, identificando los individuos fuera de los límites de control para un intervalo de confianza del 95%. Posteriormente se llevó a cabo el pre-procesado, consistente en la aplicación de un filtrado mediante el algoritmo de Savitzky-Golay con derivación, el podado de los espectros (eliminando las primeras 5 y las últimas 68 longitudes de onda) se llevó a cabo para eliminar la primera parte del espectro, afectada por el algoritmo de suavizado y la última parte del espectro, más ruidosa. En un estudio anterior (Moya-González, A. et al. 2010) se identificaron las longitudes de onda que producían una mayor aportación a los residuos para un espacio de componentes principales (Fig2) correspondiendo a la parte final de los espectros. Estas longitudes de onda no están recogidas en el MLR empleado para la estimación del SSC, por lo que su eliminación no debería suponer una pérdida de información de interés y sí una reducción del ruido presente en los espectros. Los espectros podados, que comprenden desde los 910 nm hasta los 1427 nm fueron posteriormente corregidos mediante la aplicación de la transformación de la varianza normal estándar (SNV).

La figura 3 (Fig3) muestra el proceso de tratamiento y selección de espectros llevado a cabo para la generación y validación de los distintos modelos PLS.

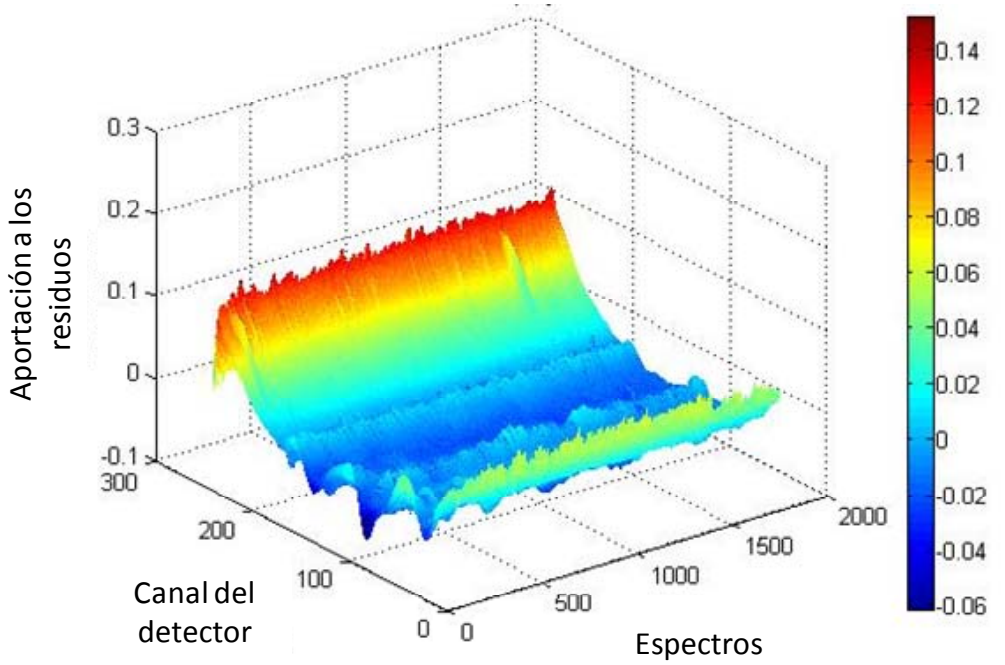

Figura 2. *Aportación al valor residual por longitudes de onda para espectros NIR de bulbos de cebolla*

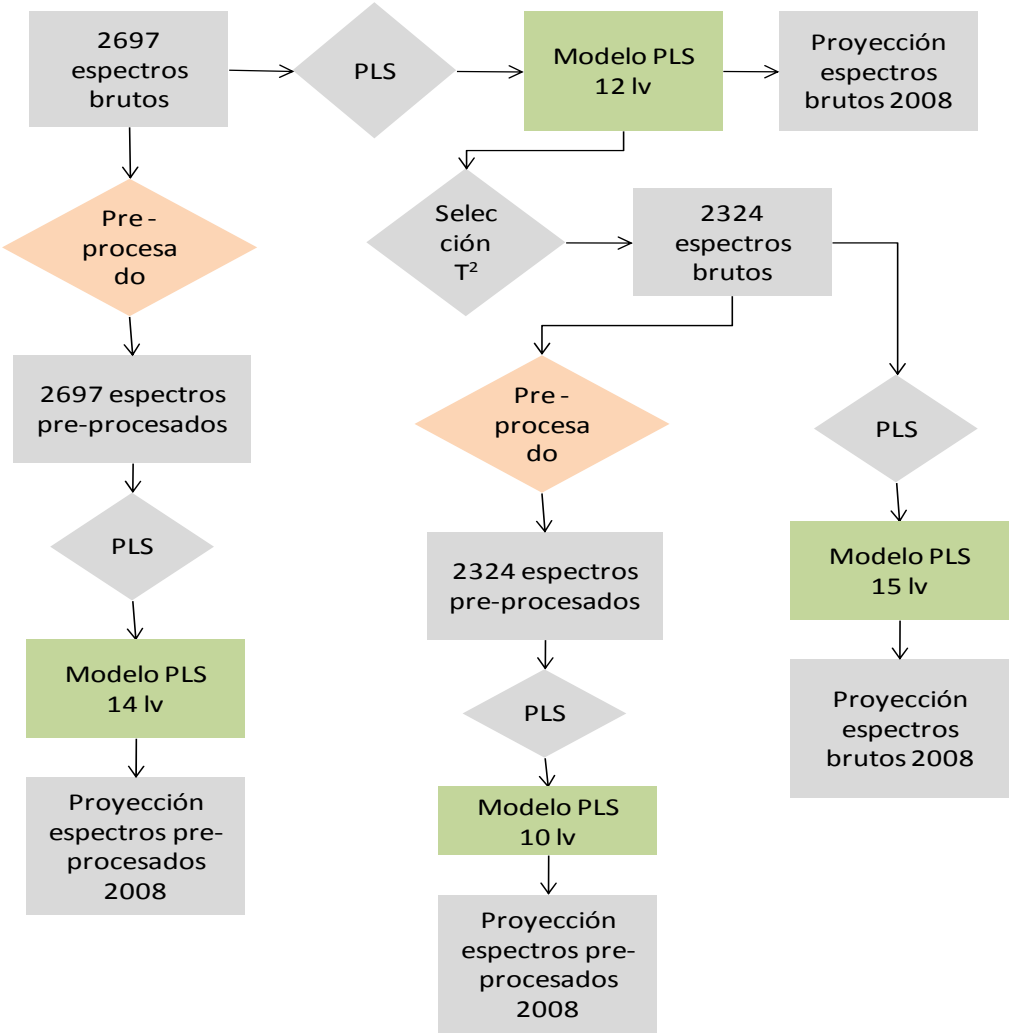

Figura 3. *Diagrama de flujo del proceso de modelización PLS, pre-procesado y selección de espectros mediante el estadístico T2 basado en modelo PLS*

3. RESULTADOS Y DISCUSIÓN

Para optimizar el número de variables latentes a emplear por el modelo PLS se ha llevado a cabo una división de los datos originales en bloques de 50 espectros. Ajustando los modelos PLS correspondientes sobre cada bloque de datos y validando los resultados obtenidos sobre bloques contiguos obtenemos el error típico de calibración (SEC) y el error típico de la estimación (SEP). Si observamos la evolución de éstos podemos comprobar cómo a medida que se incrementa el número de variables latentes, el SEC se va reduciendo, mientras que, debido al efecto de sobre-aprendizaje, llega un momento en el que el SEP comienza a crecer tras alcanzar un mínimo. Este punto nos marca el número de variables latentes que el modelo PLS debería tomar en cuenta. En el caso del modelo PLS llevado a cabo a partir de los datos de todos los espectros no procesados el número óptimo de variables latentes es de 12 (Fig 4). De forma análoga se determinó el número óptimo de variables latentes para cada uno de los modelos PLS (Tabla 1). La figura 3 (Fig 5) muestra el SSC estimado por el modelo PLS llevado a cabo a partir de los espectros sin pre-procesar, frente al valor de los SSC de referencia. En ésta, las observaciones correspondientes a espectros dentro de los límites de control para el estadístico T^2 de Hotelling aparecen representadas en azul, mientras que las correspondientes a observaciones cuya T^2 de Hotelling excede los límites de control aparecen en rojo.

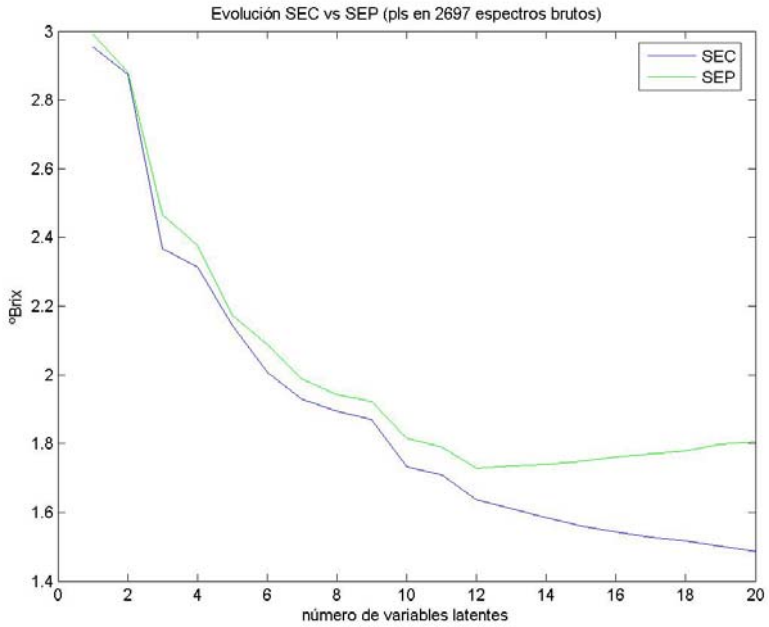

Figura 4. *Evolución de los errores típicos de calibración (SEC) y estimación (SEP) para modelos PLS ajustados sobre los espectros originales no seleccionados*

Tal y como refleja la Tabla 1, el ajuste de los modelos sobre los datos de partida no experimenta mejoras en cuanto a los errores de estimación (SEP) obtenidos, ni en cuanto al coeficiente de determinación (R2) que permanecen bastante estables independientemente del pre-procesado de los espectro o de la selección de espectros mediante la T2 de Hotelling. Las mejoras más reseñables en cuanto a los modelos PLS se reflejan en la reducción del número de variables latentes empleadas y en los parámetros obtenidos cuando los modelos son aplicados al set de datos de validación (campaña 2008). La validación del modelo ajustado sobre todos los espectros originales, con o sin selección de espectros dentro de control para la T2 de Hotelling, ofrece resultados más pobres a nivel de R2 que las validaciones llevadas a cabo de los modelos calibrados a partir de los espectros pre-procesados ((Fig 6) y (Tabla 1).

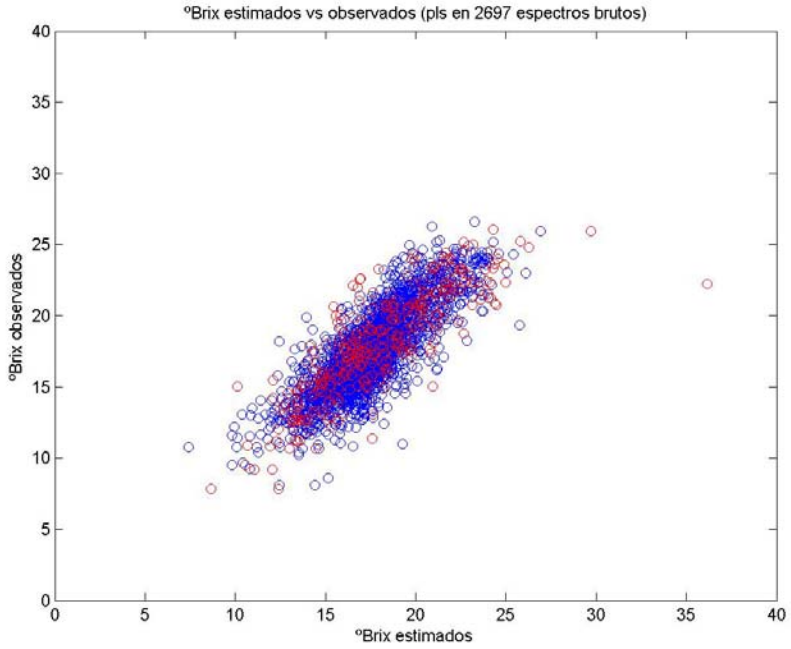

Figura 5. *SSC estimados vs SSC medidos para el modelo PLS calibrado sobre todos los espectros sin pre-procesad. Las observaciones con valores de T^2 fuera de los límites de control se representan en rojo, las correspondientes a valores de T^2 dentro de control en azul.*

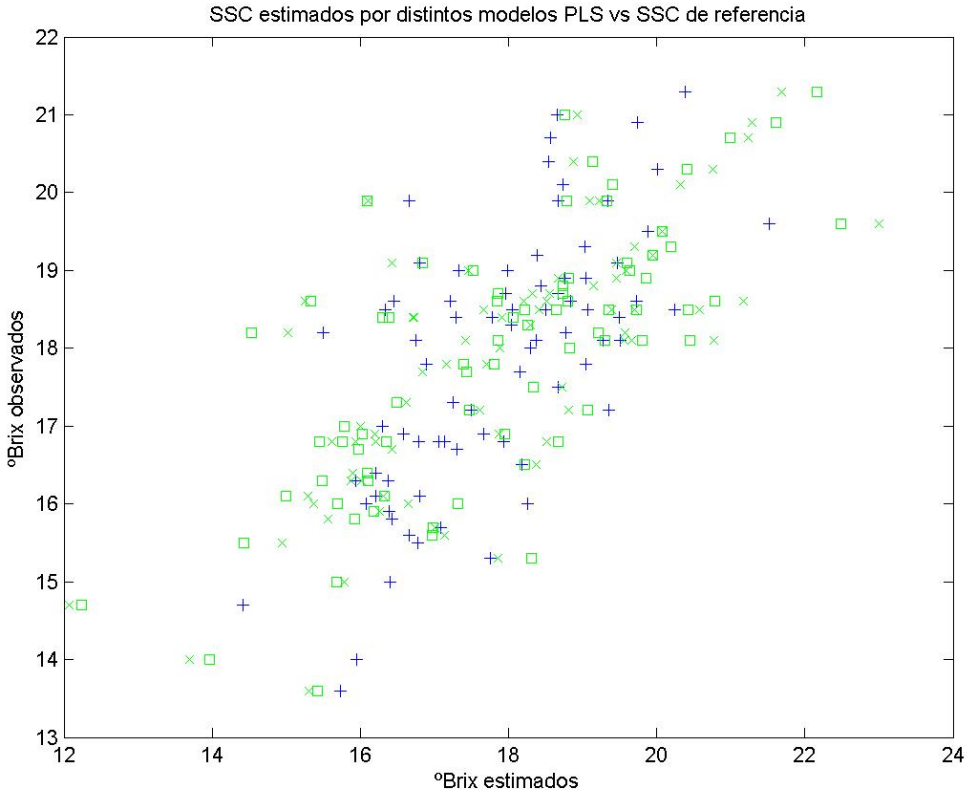

Figura 6. *SSC estimados vs SSC medidos para los espectros de la campaña 2008 sobre varios modelos PLS (+ azules para el modelo PLS generado a partir de todos los espectros brutos, x verdes para el modelo PLS generado a partir de todos los espectros pre-procesados y □ verdes para el modelo PLS generado a partir de los espectros seleccionados según su valor de T^2 y pre-procesados*

La reducción en el número de variables latentes empleado por los modelos PLS redunda en una mayor estabilidad frente a nuevos datos no incluidos en la base de datos

de calibración. El modelo desarrollado a partir de los espectros seleccionados según sus valores para la T^2 de Hotelling y pre-procesados resulta superior a los restantes ya que presenta una mejor capacidad de estimación para nuevos datos no incluidos en la base de datos de calibración. Todas la validaciones presentan SEP similares mientras que las llevadas a cabo para los espectros pre-procesados ofrecen una mayor R^2. La selección de espectros mediante sus valores para la T^2 de Hotelling permite una reducción en el número de variables latentes para los espectros pre-procesados que resulta de interés al ofrecer modelos más estables.

Tabla 1. *Parámetros del ajuste mediante modelos PLS para los distintos tipos de tratamientos empleados. Coeficiente de determinación (R^2), Error típico de la estimación (SEP) y número de variables latentes empleado por el modelo (lv)*

			Todos los espectros	**Eliminando espectros fuera de los límites de control**
Espectros empleados			2697	2324
Espectros sin pre-procesar	Datos de calibración (2001-2004)	R^2	0.66	0.67
		SEP	1.73	1.64
		lv	12	15
	Datos de validación (76 espectros) (2008)	R^2	0.45	0.42
		SEP	1.27	1.32
Espectros pre-procesados	Datos de calibración (2001-2004)	R^2	0.66	0.66
		SEP	1.73	1.66
		lv	14	10
	Datos de validación (76 espectros) (2008)	R^2	0.54	0.52
		SEP	1.39	1.4

Otro de los resultados que propicia el pre-procesado de los datos es la total desaparición de individuos fuera de control para el estadístico T^2 por lo que la mejora de los espectros obtenida mediante el pre-procesado queda bien recogida en este estadístico multivariante de control de procesos.

4. CONCLUSIONES

La combinación de las técnicas de pre-procesado de espectros produce mejoras significativas en los modelos PLS. Estas mejoras se reflejan especialmente en la validación de los modelos con datos no contemplados en la calibración.

La aplicación de estadísticos MSPC como es la T^2 de Hotelling contribuye a la mejora de los modelos PLS desarrollados sobre los espectros pre-procesados mediante una reducción en el número de variables latentes empleado. Este estadístico también recoge las mejoras obtenidas mediante el pre-procesado ya que para los espectros pre-procesados en ningún caso sus valores superan los límites de control.

El filtrado y pre-procesado de los espectros, combinado con la aplicación de estadísticos MSPC permite mejorar la capacidad de estimación de los modelos PLS pero un mejor conocimiento de las fuentes de variación interferentes que intervienen en la determinación de calidad de un material biológico en evolución, como son las cebollas, y en un entorno industrial cambiante, resultará necesario para lograr futuros avances.

5. AGRADECIMIENTOS

Los autores agradecen la financiación del presente trabajo a la Universidad Politécnica de Madrid a través del proyecto DURASFRUT II (AL11-P(I+D)-06)

6. BIBLIOGRAFIA

Barreiro, P., L. Henche, et al. (2004). *Multivariate diagnosis of the variability of NIR spectrometers under industrial applications. SJAR* 2(4): 485-492.

Barreiro, P., F. Chauchard, et al. (2005). *Robust modelling for at-line and on-line calibration transfer in a NIR industrial application. Chemiometrie*. Lille, France.

Barreiro, P., A. Moya-González, et al. (2005). *Analysis of the effect of product temperature on the segregation of onions by means of online NIR spectrometry. FRUTIC 05, Information and technology for sustainable fruit and vegetable production. 7th Fruit nut and vegetable production engineering symposium*. Cemagref. Montpellier, France: 473 - 482.

Barreiro, P., M. Ruiz-Altisent, et al. (2005). *Multivariate analysis of an on-line NIR spectrometer under industrial use. Proceedings of the 3rd International Symposium on Applications of Modelling as an Innovative Technology in the Agri-Food Chain*(674): 513-519.

Moya-González, A., P. Barreiro, et al. (2006). *Calibration transfer techniques for on-line NIR evaluation of SSC in onions. VI CIGR World Congress. Agricultural engineering for a better world*. Bonn (Germany): Book of abstracts. pp: 585-586.

Moya-González, A., P. Barreiro, et al. (2009). *Diagnóstico de la operación de un Espectrómetro NIR montado en línea mediante Análisis Multivariante. V Congreso Nacional y lll Congreso lbérico Agrolngeniería 2009*. Lugo, Spain.

Moya-González, A., P. Barreiro, et al. (2010). *Procedure for calibration transfer between seasons for on-line NIR evaluation of SSC in onion breeding lines. International Conference on Agricultural Engineering*. Clermont-Ferrand (France).

Ortiz-Cañavate, J., A. Moya-González, et al. (2010). *Identification and classification of out of control measurements of a NIR spectrometer under industrial use for onion quality determination. 17th CIGR World Congress*. Québec City, Canada.

Zeaiter, M., D. Rutledge, et al. (2009). *Preprocessing Methods. Comprehensive Chemometrics*. Oxford, Elsevier: 121-231.